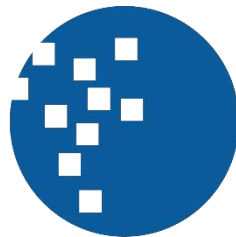


**ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI  
MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID  
VADER, TF-IDF, BERT DAN BERTOPIC**



**UMN**

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

**SKRIPSI**

**Cindy Febriani Santoso**

**00000059735**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA  
TANGERANG**

**2025**

**ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI  
MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID  
VADER, TF-IDF, BERT DAN BERTOPIC**



Diajukan sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana Komputer

**Cindy Febriani Santoso**

**0000059735**

**PROGRAM STUDI SISTEM INFORMASI  
FAKULTAS TEKNIK DAN INFORMATIKA  
UNIVERSITAS MULTIMEDIA NUSANTARA  
TANGERANG**

**2025**

i

## HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Cindy Febriani Santoso

Nomor Induk Mahasiswa : 00000059735

Program Studi : Sistem Informasi

Skripsi dengan judul:

**ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI  
MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID  
VADER, TF-IDF, BERT DAN BERTOPIC**

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 25 Mei 2025

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



(Cindy Febriani Santoso)

## HALAMAN PERSETUJUAN

Skripsi dengan judul

### **ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID VADER, TF-IDF, BERT DAN BERTOPIC**

Oleh

Nama : Cindy Febriani Santoso

NIM : 00000059735

Program Studi : Sistem Informasi

Fakultas : Teknik dan Informatika

Telah disetujui untuk diajukan pada

Sidang Ujian Skripsi Universitas Multimedia Nusantara

Tangerang, 4 Juni 2025

Pembimbing



Ahmad Faza, S.Kom., M.T.I.

Ketua Program Studi Sistem Informasi



Ririn Ikana Desanti, S.Kom., M.Kom.

## HALAMAN PENGESAHAN

Skripsi dengan judul


### **ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID VADER, TF-IDF, BERT DAN BERTOPIC**

Oleh

Nama : Cindy Febriani Santoso  
NIM : 00000059735  
Program Studi : Sistem Informasi  
Fakultas : Teknik dan Informatika

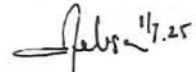
Telah diujikan pada Jumat, 20 Juni 2025  
Pukul 15.00 s.d 17.00 dan dinyatakan  
**LULUS**

Dengan susunan penguji sebagai berikut.

Ketua Sidang  
  
01/07  
2025


Jansen Wiratama, S.Kom., M.Kom.  
0409019301

Penguji

  
17.25

Melissa Indah Fianty, S.Kom., MMSI.  
0313019201

Pembimbing

  
1/7/2025

Ahmad Faza, S.Kom., M.T.I.  
0312019501

Ketua Program Studi Sistem Informasi

  
1/7/25

Ririn Ikana Desanti, S.Kom., M.Kom.  
0313058001

iv

## HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama : Cindy Febriani Santoso

NIM : 00000059735

Program Studi : Sistem Informasi

Jenjang : S1

Judul Karya Ilmiah :

**ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI  
MENGENAI CORETAX MENGGUNAKAN MODEL  
HYBRID VADER, TF-IDF, BERT DAN BERTOPIC**

Menyatakan dengan sesungguhnya bahwa saya bersedia\* (pilih salah satu):

- Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) \*\*.
- Lainnya, pilih salah satu:
  - Hanya dapat diakses secara internal Universitas Multimedia Nusantara
  - Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 25 Mei 2025



(Cindy Febriani Santoso)

## KATA PENGANTAR

Puji Syukur penulis sampaikan atas terselesainya penelitian dan penulisan skripsi ini dengan judul, “Analisis Sentimen dan Pemodelan Topik Opini mengenai Coretax menggunakan Model Hybrid VADER, TF-IDF, BERT, dan BERTopic.” Pembuatan skripsi ini dilakukan untuk memenuhi salah satu syarat mendapat gelar Sarjana Komputer Program Studi Sistem Informasi pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Penulis menyadari bahwa, tanpa bantuan, dukungan, dan bimbingan berbagai pihak sejak memulai masa perkuliahan hingga penyusunan skripsi ini, sangatlah sulit bagi penulis untuk menyelesaikan skripsi ini. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Bapak Dr. Andrey Andoko, selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti, S.Kom., M.Kom., selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara.
4. Bapak Jansen Wiratama, S.Kom., M.Kom., selaku Ketua Sidang yang memberikan kritik dan saran membangun selama proses sidang berlangsung, sehingga penulisan skripsi ini menjadi lebih baik.
5. Ibu Melissa Indah Fianty, S.Kom., MMSI., selaku Penguji yang memberikan kritik dan saran membangun selama proses sidang berlangsung, sehingga penulisan skripsi ini menjadi lebih baik.
6. Bapak Ahmad Faza, S.Kom., M.T.I., sebagai Pembimbing yang telah meluangkan waktu untuk memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
7. Kak Desy dan Mas Ryan yang meluangkan waktunya sebagai narasumber untuk melakukan validasi hasil temuan yang didapat pada skripsi ini.
8. Keluarga penulis sebagai sumber kekuatan dan motivasi yang juga memberikan dukungan secara material, mental, dan moral yang menjadi pendorong utama bagi penulis sehingga dapat menyelesaikan skripsi ini.

9. Teman-teman yang hadir dan mendampingi penulis dalam seluruh dinamika selama menjalani masa perkuliahan dari awal hingga terselesaikannya skripsi ini.

Semoga skripsi ini dapat bermanfaat bagi pembaca dalam memahami respons publik terhadap layanan perpajakan yang diluncurkan pemerintah, berkontribusi pada ilmu berkaitan dengan *Natural Language Processing*, serta menjadi referensi untuk penelitian selanjutnya di masa depan. Penulis juga berharap hasil penelitian dalam skripsi ini dapat memberikan wawasan dan dampak positif, terutama dalam meningkatkan literasi terhadap sistem perpajakan di dunia nyata.

Tangerang, 25 Mei 2025



(Cindy Febriani Santoso)

UMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



# ANALISIS SENTIMEN DAN PEMODELAN TOPIK OPINI MENGENAI CORETAX MENGGUNAKAN MODEL HYBRID VADER, TF-IDF, BERT DAN BERTOPIC

(Cindy Febriani Santoso)

## ABSTRAK

Coretax merupakan sistem perpajakan yang diluncurkan oleh pemerintah Indonesia pada bulan Januari 2025. Sistem ini mengintegrasikan seluruh layanan DJP online ke dalam satu platform. Implementasi sistem Coretax diharapkan dapat meningkatkan penerimaan pajak akibat kepatuhan pajak yang rendah. Sejak perilisannya, sistem ini menghasilkan opini dan reaksi publik di media sosial, terutama X. Penelitian ini melakukan analisis sentimen dan pemodelan topik untuk mengetahui opini publik terhadap Coretax.

Metode analisis sentimen menggabungkan fitur VADER, TF-IDF, dan BERT embeddings ke model klasifikasi Logistic Regression. Sementara itu, pemodelan topik diimplementasikan dengan BERTopic untuk mengekstrak topik yang dibicarakan di X. Penelitian ini menerapkan KDD sebagai kerangka kerja penelitian. Proses pembersihan data terdiri dari penyaringan dan pengubahan teks menjadi huruf kecil, penerjemahan ke bahasa Inggris, tokenisasi, lemmatisasi, dan pemeriksaan ejaan. Di sisi lain, transformasi data dilakukan dengan pelabelan data dan SMOTE untuk menyeimbangkan kelas data. Tahap-tahap sebelumnya menghasilkan 17.100 data valid untuk analisis lebih lanjut.

Penelitian ini menunjukkan bahwa model *hybrid* dapat mengungkap dominasi sentimen negatif terhadap Coretax di Indonesia pada data aktual. Metode penanganan kelas tidak seimbang SMOTE memiliki performa yang tinggi dengan akurasi, presisi, *recall*, dan *f1-score* 94%. BERTopic dapat mengidentifikasi apresiasi fitur yang didiskusikan secara positif dan keluhan teknis yang dihadapi pengguna pada sentimen negatif. Validasi hasil mengkonfirmasi relevansi hasil temuan dan analisis kebijakan lebih lanjut menunjukkan inkonsistensi dengan Permenkomdigi Nomor 6 Tahun 2025. Sistem Coretax yang tidak optimal berkontribusi pada penurunan penerimaan pajak sehingga dibutuhkan perbaikan terhadap keluhan pengguna.

**Kata kunci:** Coretax, sentimen, opini, topik, X

**SENTIMENT ANALYSIS AND TOPIC MODELING OF PUBLIC  
OPINION ON CORETAX USING VADER, TF-IDF, BERT  
HYBRID MODEL, AND BERTOPIC**

(Cindy Febriani Santoso)

**ABSTRACT (English)**

*Newly launched Indonesian government online tax system, Coretax, have just launched in January 2025. It integrates all services in DJP online into a single platform. Coretax is expected to increase tax revenue due to low tax compliance. Since the release date, this new systems led to public opinions and reactions in social media, especially X. This research conducts sentiment analysis and topic modeling to find out public opinion on Coretax implementation.*

*Sentiment analysis method integrates VADER, TF-IDF, and BERT embeddings features into Logistic Regression model. As for topic modeling implemented using BERTopic to extract what is being discussed in X. This study employs KDD framework as flow of research. Data cleaning include text filtering and case folding, English translation, tokenization, lemmatization, and spellchecking. On the other hand, data transformation is done with data labeling and SMOTE class balancing. These steps resulted in 17.100 tweets for further analysis.*

*This research shows that proposed hybrid model can reveal the dominance of negative sentiment towards Coretax in Indonesia on actual data. While using SMOTE, the model performs highly with 94% accuracy, precision, recall, and f1-score. BERTopic is able to identify discussion of appreciation positively and technical complaints by users from negative sentiment. Expert validation confirms the relevance of findings. Further policy analysis shows that Coretax does not comply with Permenkomdigi Number 6 of 2025. Inefficiency caused by Coretax has led to decline in tax revenue, making it necessary to address user complaints.*

**Kata kunci:** *Coretax, opinion, sentiment, topic, X*

## DAFTAR ISI

HALAMAN PERNYATAAN TIDAK PLAGIAT .....	ii
HALAMAN PERSETUJUAN .....	iii
HALAMAN PENGESAHAN .....	iv
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH .....	v
KATA PENGANTAR.....	vi
ABSTRAK .....	viii
<i>ABSTRACT (English)</i> .....	ix
DAFTAR TABEL .....	xiii
DAFTAR GAMBAR.....	xiv
DAFTAR RUMUS .....	xv
DAFTAR LAMPIRAN .....	xvi
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
<b>1.1 Latar Belakang.....</b>	<b>1</b>
<b>1.2 Rumusan Masalah.....</b>	<b>3</b>
<b>1.3 Batasan Masalah .....</b>	<b>4</b>
<b>1.4 Tujuan dan Manfaat Penelitian.....</b>	<b>4</b>
<b>1.4.1 Tujuan Penelitian.....</b>	<b>4</b>
<b>1.4.2 Manfaat Penelitian.....</b>	<b>4</b>
<b>1.5 Sistematika Penulisan .....</b>	<b>5</b>
<b>BAB II LANDASAN TEORI .....</b>	<b>7</b>
<b>2.1 Penelitian Terkait.....</b>	<b>7</b>
<b>2.2 Teori Penelitian .....</b>	<b>11</b>
<b>2.2.1 Analisis sentimen.....</b>	<b>11</b>
<b>2.2.2 Pemodelan Topik.....</b>	<b>11</b>
<b>2.2.3 Pre-processing .....</b>	<b>11</b>
<b>2.3 Framework dan Algoritma Penelitian .....</b>	<b>13</b>
<b>2.3.1 KDD.....</b>	<b>13</b>
<b>2.3.2 TF-IDF .....</b>	<b>13</b>
<b>2.3.3 SMOTE .....</b>	<b>14</b>
<b>2.3.4 BERT.....</b>	<b>15</b>

2.3.5	DistilBERT.....	15
2.3.6	VADER.....	16
2.3.7	Logistic Regression.....	16
2.3.8	BERTopic.....	17
2.3.9	Confusion Matrix.....	17
2.3.10	Coherence score.....	19
2.4	Tools Penelitian.....	20
2.4.1	Google Colab.....	20
<b>BAB III METODOLOGI PENELITIAN.....</b>		<b>22</b>
3.1	Gambaran Umum Objek Penelitian.....	22
3.2	Metode Penelitian.....	23
3.2.1	Alur Penelitian.....	23
3.3	Variabel Penelitian.....	28
3.3.1	Analisis Sentimen.....	28
3.3.2	Pemodelan topik.....	29
3.4	Teknik Pengumpulan Data.....	29
3.4.1	Tweet Harvest.....	29
3.5	Teknik Analisis Data.....	29
<b>BAB IV ANALISIS DAN HASIL PENELITIAN.....</b>		<b>31</b>
4.1	Data selection.....	31
4.2	Data Cleaning.....	33
4.2.1	Text filtering dan Case Folding.....	33
4.2.2	Translate data.....	34
4.2.3	Tokenization.....	35
4.2.4	Lemmatization.....	36
4.2.5	Spellchecking.....	36
4.2.6	Remove stopwords.....	37
4.3	Data Transformation.....	38
4.3.1	Data labeling otomatis.....	38
4.3.2	SMOTE.....	39
4.4	Data Mining.....	39
4.4.1	Analisis Sentimen.....	39

4.4.2	Pemodelan Topik.....	43
4.5	Interpretation/Evaluation .....	46
4.5.1	Confusion Matrix .....	46
4.5.2	Coherence Score .....	57
4.6	Pembahasan .....	61
<b>BAB V</b>	<b>SIMPULAN DAN SARAN.....</b>	<b>68</b>
5.1	Simpulan .....	68
5.2	Limitasi .....	68
5.3	Saran .....	69
<b>DAFTAR PUSTAKA.....</b>		<b>70</b>
<b>LAMPIRAN.....</b>		<b>76</b>



## DAFTAR TABEL

Tabel 2.1 Penelitian terkait dalam lima tahun terakhir .....	7
Tabel 2.2 Perbandingan framework CRISP-DM, SEMMA, dan KDD .....	13
Tabel 2.3 Contoh <i>valence score</i> dalam leksikon VADER.....	16
Tabel 2.4 Tabel Confusion Matrix .....	18
Tabel 3.1 Deskripsi atribut dataset objek penelitian .....	22
Tabel 3.2 Hasil data selection .....	24
Tabel 3.3 Parameter dalam model DistilBERT .....	26
Tabel 4.1 Contoh data tweets pada atribut <code>full_text</code> .....	32
Tabel 4.2 Tweets setelah melalui tahap filtering dan case folding .....	33
Tabel 4.3 Contoh tweets setelah diterjemahkan.....	34
Tabel 4.4 Contoh tweets setelah tokenization.....	35
Tabel 4.5 Contoh tweets setelah lemmatization.....	36
Tabel 4.6 Contoh tweets setelah spellchecking dengan library hunspell.....	36
Tabel 4.7 Contoh tweets setelah dilakukan remove stopwords .....	37
Tabel 4.8 Jumlah kelas sentimen dengan label otomatis .....	38
Tabel 4.9 Jumlah data training dan testing .....	43
Tabel 4.10 Parameter pemodelan topik BERTopic.....	43
Tabel 4.11 Performa masing-masing model dan rasio splitting data.....	61
Tabel 4.12 Perbandingan performa analisis sentimen dengan penelitian terkait ..	62
Tabel 4.13 Perbandingan jumlah data valid dan tahap preprocessing .....	63
Tabel 4.14 Perbandingan performa pemodelan topik dengan penelitian terkait...	66



## DAFTAR GAMBAR

Gambar 3.1 Alur penelitian.....	23
Gambar 3.2 Alur data cleaning .....	24
Gambar 3.3 Alur analisis sentimen dan pemodelan topik dalam data mining.....	27
Gambar 4.1 Hasil output sentimen VADER.....	40
Gambar 4.2 Distribusi sentimen dengan leksikon VADER.....	40
Gambar 4.3 Terms dengan skor TF-IDF tertinggi .....	41
Gambar 4.4 Visualisasi hubungan antar topik pada sentimen positif.....	44
Gambar 4.5 Visualisasi hubungan antar topik pada sentimen negatif.....	45
Gambar 4.6 Metrik klasifikasi VADER tanpa SMOTE.....	46
Gambar 4.7 Metrik klasifikasi VADER dengan SMOTE.....	47
Gambar 4.8 Metrik VADER dan TF-IDF 60:40.....	48
Gambar 4.9 Metrik VADER dan TF-IDF 70:30.....	48
Gambar 4.10 Metrik VADER dan TF-IDF 80:20.....	49
Gambar 4.11 Metrik VADER dan TF-IDF 90:10.....	49
Gambar 4.12 Metrik VADER dan TF-IDF dengan SMOTE 60:40.....	50
Gambar 4.13 Metrik VADER dan TF-IDF dengan SMOTE 70:30.....	50
Gambar 4.14 Metrik VADER dan TF-IDF dengan SMOTE 80:20.....	51
Gambar 4.15 Metrik VADER dan TF-IDF dengan SMOTE 90:10.....	51
Gambar 4.16 Metrik VADER, TF-IDF, dan BERT 60:40.....	52
Gambar 4.17 Metrik VADER, TF-IDF, dan BERT 70:30.....	53
Gambar 4.18 Metrik VADER, TF-IDF, dan BERT 80:20.....	53
Gambar 4.19 Metrik VADER, TF-IDF, dan BERT 90:10.....	54
Gambar 4.20 Metrik VADER, TF-IDF, dan BERT dengan SMOTE 60:40 .....	55
Gambar 4.21 Metrik VADER, TF-IDF, dan BERT dengan SMOTE 70:30 .....	55
Gambar 4.22 Metrik VADER, TF-IDF, dan BERT dengan SMOTE 80:20 .....	56
Gambar 4.23 Metrik VADER, TF-IDF, dan BERT dengan SMOTE 90:10 .....	56
Gambar 4.24 C <sub>v</sub> coherence score topik sentimen positif.....	57
Gambar 4.25 Wordcloud sentimen positif topik 0 .....	57
Gambar 4.26 Wordcloud sentimen positif topik 3 .....	58
Gambar 4.27 Wordcloud sentimen positif topik 2 .....	58
Gambar 4.28 C <sub>v</sub> coherence score topik sentimen negatif.....	59
Gambar 4.29 Wordcloud sentimen negatif topik 1 .....	59
Gambar 4.30 Wordcloud sentimen negatif topik 2 .....	60
Gambar 4.31 Wordcloud sentimen negatif topik 5 .....	61
Gambar 4.32 Hasil prediksi sentimen dengan model hybrid .....	64

## DAFTAR RUMUS

Rumus 2.1 Rumus term frequency .....	14
Rumus 2.2 Rumus inverse document frequency .....	14
Rumus 2.3 Rumus TF-IDF .....	14
Rumus 2.4 Rumus sigmoid .....	17
Rumus 2.5 Rumus metrik accuracy .....	18
Rumus 2.6 Rumus recall .....	18
Rumus 2.7 Rumus precision .....	19
Rumus 2.8 Rumus f1-score .....	19
Rumus 2.9 Rumus segmentasi kata coherence score .....	19
Rumus 2.10 Rumus pembuatan vektor konteks data .....	19
Rumus 2.11 Rumus cosine similarity coherence score .....	20
Rumus 2.12 Rumus agregasi coherence score .....	20





## DAFTAR LAMPIRAN

Lampiran A Turnitin Similarity Report .....	76
Lampiran B Form Konsultasi Bimbingan .....	82
Lampiran C Kode <i>import library</i> untuk <i>scrapping data</i> .....	83
Lampiran D Kode <i>scrapping data</i> .....	83
Lampiran E Kode penggabungan seluruh dataset .....	83
Lampiran F Implementasi filtering data .....	83
Lampiran G Fungsi <i>text filtering</i> .....	84
Lampiran H Penerjemahan dengan GoogleTranslator .....	84
Lampiran I Kode implementasi <i>tokenization</i> .....	84
Lampiran J Kode implementasi <i>lemmatization</i> dengan WordNet .....	85
Lampiran K Kode fungsi <i>spellchecking</i> dengan Hunspell .....	85
Lampiran L Kode <i>remove stopwords</i> pada dataset .....	85
Lampiran M Kode <i>sentiment labeling</i> dengan DistilBERT .....	86
Lampiran N Kode implementasi VADER .....	86
Lampiran O Kode <i>labeling polarity score</i> VADER .....	86
Lampiran P Kode deklarasi fungsi BERT .....	86
Lampiran Q Kode implementasi Logistic Regression .....	87
Lampiran R Kode BERTopic untuk sentimen positif .....	87
Lampiran S Kode BERTopic untuk sentimen negatif .....	87
Lampiran T Transkrip Wawancara dengan Narasumber (Kak Desy) .....	88
Lampiran U Transkrip Wawancara dengan Narasumber (Mas Ryan) .....	89

